Response to Arizona Department of Education RFI for
School accountability system and components.

RFI due July 15th, 2016 3pm MST

The National Center for the Improvement of Educational Assessment, Inc. (Center for Assessment), a Dover, NH based not-for-profit corporation is pleased to provide responses, feedback, suggestions and comments to items requested by the Arizona Department of Education (ADE) in the Request for Information (RFI) - RFI# ADED16-0002.

The RFI outlines information on various topics and offers the opportunity for entities to respond to one, any, or all of the questions contained in the RFI. As indicated: "This RFI solicits feedback from interested parties with an relevant expertise, systems, or methodology they have developed or conceptualized which meet the intent of any one or more of the components described below".

The components of information being requested are as follows. The Center for Assessment's response to each of the components is provided after each of the five numbered components.

1. Demonstration of values

a. How does a transparent and fair accountability system define an "excellent" school in regards to:
   i. Preparing all students for College/Career readiness
   ii. Improving achievement and outcomes among student subgroups
   iii. Graduating students prepared for postsecondary workforce and/or education
   iv. Demonstrating growth on standardized assessments aligned to Arizona's standards
   v. Providing a high-quality, well-rounded education to families regardless of income
   vi. Meeting the needs of parents and students in the community

b. How does a transparent and fair accountability system define a "failing" school in regards to:
   i. Preparing all students for College/Career readiness
   ii. Improving achievement and outcomes among student subgroups
   iii. Graduating students prepared for postsecondary workforce and/or education
   iv. Demonstrating growth on standardized assessments aligned to Arizona's standards
   v. Providing a high-quality, well-rounded education to families regardless of income
   vi. Meeting the needs of parents and students in the community

c. How does a transparent and fair accountability system differentiate among schools that are neither "excellent" nor "failing"?

Response from the Center for Assessment:

In responding to this component of the RFI, the Center for Assessment will briefly offer the context of accountability and an overview of the federal requirements for accountability and then provide information in response to this component. Even though the component is organized in three parts, it seems the three parts are asking how the transparent and fair accountability system differentiates among schools that are categorized in one of three levels of "excellent", "failing", or neither. So, the response provided will address all three aspects together and offer some insights into answering the question posed.

The development of a school accountability system involves a process of documenting and reflecting the values of the educational system while complying with state and federal policies. A collective of states, wanting to transform accountability systems while waiting for reauthorization of the NCLB, made the following recommended goals (CCSSO, 2011):

1. Readiness for college, career, and civic responsibilities through the mastery of rigorous content knowledge and successful application of knowledge using higher-order skills and dispositions.
2. Accountability systems should promote school and district performance improvement and student achievement and growth by providing timely, actionable information.
3. Accountability systems should offer incentives for academic achievement for all students.
4. Accountability systems should encourage and not impeded personalized or performance-based teaching and learning including support and targeted information to build capacity to help leaders and educators improve.

The Every Student Succeeds Act (ESSA) is the latest reauthorization of the Elementary and Secondary Education Act of 1965. It represents an omnibus program comprised of nine major titles with assessment and accountability falling under Title I. Title I provides approximately $60 billion for the education of disadvantaged students. Waivers will expire in the summer of 2016 with accountability transition to occur during the 2016-2017 school year and going live in 2017-2018. Standards and assessment peer review are required having started in the early part of 2016.

The required state plan must have the following components:

1. State standards
2. Academic assessments
3. Statewide accountability and reporting system
4. Approach to school improvement and support

5. Indications of how the state will support evidence-based distribution programs with fiscal flexibility and transparency

The school accountability determinations are state determined within some parameters. Under ESSA, states can establish goals for status and improvement for (a) academic achievement, (b) graduation rates, and (c) the sub-groups' performance. More specifically the five types of indicators required are as follows:

1. Academic Achievement (e.g., proficiency)
2. Another valid and reliable academic indicator (e.g., growth gap closure)
3. Graduation rates (specifically Adjusted Cohort Graduation Rate)
4. English language proficiency
5. Indicator of school quality or success that meaningfully differentiates and is valid, reliable, and comparable.

The law indicates that much greater weight must be given to the first four indicators, but regulations and guidance have not yet been provided to offer more specificity. It is likely that one could operationalize this for the weight for the fifth indicator as ranging from 1% to 40%.

Using the combination of these indicators, the state must specify the indicators and methodology that starting in 2017-2018 and at least once every three years afterwards produces a statewide category of schools for comprehensive support and improvement for schools using the following categories:

- Lowest performing 5% of schools
- High school graduation rate less than 67%
- School have low performing sub-groups

States can increase the frequency of these determinations and include more performance categories. According to the draft regulations (as of May 31, 2016), states are required to report assessment results annually, disaggregated by the federal accountability and reporting subgroups. A summative rating, which the methodology is determined by the states using the five indicators, must be produced and reported annually. After the calculation of the summative rating, the lowest performing schools will be identified annually for target support based on sub-group performance.

First, the manner in which the accountability system is designed in order to reflect the values of Arizona and to ensure a transparent and fair system is to involve key stakeholders and secure meaningful input. While accountability systems cannot be designed by hundreds of people, a systematic process for undertaking this involves a number of steps including the engagement of key stakeholders. This process of engaging the stakeholders is consistent with the goals and proposals indicated by Superintendent Douglas as documented in the *AZ Kids Can't Afford to Wait* (2015) document.

Second, in order for the accountability system to identify schools that are "excellent", "failing", or neither (as indicated in this component of the RFI) using the six stated elements involves a number of technical and procedural components. To ensure that the accountability system meets its intended goals and does so with fidelity involves three major components: (1) The specification of the goals and values of the accountability system, (2) the validity of the design of the accountability system including the indicators used, and (3) gathering evidence that the accountability system works in meeting its goals. In order to undertake these steps and maintain the values of a transparent and fair system, it is important to utilize a process that involves a leadership team, a steering committee of stakeholders, and a technical advisory committee. Each of these is briefly described below starting with the need for three sets of teams/committees.

Teams/Committees

The function of each team/committee is to ensure that the design and implementation of the accountability system effectively promotes the state's goals and policy priorities, as articulated by key leaders and stakeholders, and operationalizes these goals in a system that is technically defensible. The role and function of each team is briefly provided below.

1. Leadership Team – This represents a small number of Department of Education leaders (e.g. two or three) who have responsibility of the accountability system and related areas (e.g., assessment or data). All design and implementation functions would fall under this committee.

2. Steering Committee – This represents a moderately-sized committee (e.g. six to ten) of representatives of stakeholder groups. The purpose of this steering committee is to provide feedback from the perspective of the various stakeholder groups to the leadership team regarding design and policy elements. A second role of this committee is to take information back to the stakeholder groups that they represent as one of many channels for communication for the purpose of transparency. It is expected that the Steering Committee would meet regularly during the initial two years of the design and beginning of the implementation and then once or twice a year during the implementation. This Committee would be facilitated by someone from the Leadership Team and/or an external consultant working closely with the leadership team.

3. Technical Advisory Committee – This represents an external committee of national experts who will provide the leadership team and the Department with feedback and advice on technical aspects of the accountability system, as well as support the quality of the accountability system. The Leadership Team should be responsible for the Technical Advisory Committee. Members of this committee would represent expertise in assessment, accountability,

curriculum, learning science, and applied educational research. They would serve on this committee for up to three years and provide oral and sometimes written feedback to the Department of Education. The Technical Advisory Committee should meet approximately three or four times annually, usually for one or two days. The meeting would be preceded with an agenda and material to review at least two weeks before the meeting, such that the Technical Advisory Committee can provide the technical assistance and feedback needed. In order to maximize the efficiency and quality of the information provided, an experience facilitator or chair of the committee should be utilized. Generally, honoraria, as well as travel reimbursements are provided to these committee members.

Specification of Goals

The best accountability systems set clear goals for people. They should be meaningful, challenging, achievable, and measurable. They provide regular information and feedback to all and guide the work. Unfortunately, most accountability systems do not do these things very well (Pelzman & Domaleski, 2010). Effective accountability systems seek to improve schools by clearly indicating goals and expectations for improvement that are linked to strategies that build capacity of educators to deliver (Cour, Porter, Rome, & Towne, 2010).

The specification of these goals utilizes existing documents and statements, but requires a review and engaging stakeholders. The Leadership Team and Steering Committee members would represent the initial groups to engage in the clarification of the goals.

As part of the specification of the goals, the components of the current accountability system must be evaluated both conceptually to see the alignment with the goals and empirically to evaluate the impact of the current systems.

Validity of the Design and Indicators

As indicated ESSA specifies the use of the following indicators:

1. Academic Achievement (e.g., proficiency)
2. Another valid and reliable academic indicator (e.g., growth gap closure)
3. Graduation rates (specifically Adjusted Cohort Graduation Rate)
4. English language proficiency
5. Indicator of school quality or success that meaningfully differentiates and is valid, reliable, and comparable.

In principle, five of the six components indicated in the RFI under the first informational component of "Demonstration of Values" fit under these indicators. The efficacy of these indicators depends on how these indicators are operationalized, the methodology involved in their development, collection, and

aggregation, the extent to which they represent the goals of the accountability system, and, most importantly, the educational context in which they are embedded (Tucker, 2014).

The sixth component of meeting the needs of parents and students in the community is rather vague. But, in following the process of engaging stakeholders to (a) understand the needs and (b) incorporate indicators to represent these needs, followed by developing valid indicators with empirical evidence to support these needs will permit for the effective identification of schools.

While a perfect accountability system does not exist, there are key elements and practices (Cour, Porter, Rome, & Towne, 2010) that are relevant to Arizona:

1. Promoting College and career readiness and on-track to readiness is an appropriate outcome for the K-12 system.
2. Results must be communicated in a timely and effective manner to all stakeholders.
3. Tools and resources for schools and districts to use results must be provided.

The six elements indicated in the first component of the RFI are related to the elements in the next-generation accountability systems (CCSSO, 2011, p.12):

- Focus on a minimum, specific goal of college and career readiness upon high school graduation.
- Encourage continuous, significant student growth toward college- and career-readiness.
- Understand that what is measured and reported must be tightly linked to requisite actions, supports, and interventions to best improve student achievement.
- Annual determinations coupled with diagnostic reviews provide clear and meaningful information to drive school and district performance.
- Purposefully integrate each element of the system so that one informs the other, creating greater effectiveness and resource efficiency.
- Provide incentives for growth and achievement at all levels of performance
- Connect with and are balanced across other reforms.
- Recognize the tight locus of control between districts and their schools, and seek to build capacity within districts for supporting their schools and holding them accountable for the same.
- Give particular and meaningful focus to the lowest-performing schools and districts.
- Place the student at the center of the system by promoting high-quality instruction and reinforcing the importance of sound teaching and learning practices.
- Recognize that motivation is a strong component of success and contributes to strong and positive school cultures.

- Are dynamic – promoting continual innovation and improvement based on evaluation of the accountability system and emerging technologies.

To make the determinations of "excellent", "failing", or neither the indicators and methodology used must be valid and reliable in making meaning distinctions between and within low- and high-performing schools. Disaggregated data must be used to ensure that underperformance of any student subgroup as well as achievement gaps between subgroups are transparent. The key issues that need to be addressed are (a) the weighting, (b) rules for aggregation (e.g. compensatory/conjunctive), (c) performance standards and (d) business rules (e.g. allowance of exceptions.)

Evidence that It Works

While the design of the accountability can be done using approaches that have been demonstrated to be effective, as well as from experience with successful systems, evidence must be gathered to ensure that the identification of schools into the categories of "excellent", "failing", and neither is done in a valid manner. This validation evidence is gathered through the use of scientific methods utilized in rigorous program evaluation or applied research. This approach underlies all of the proposals expressed by Superintendent Douglas to be responsive to the needs of the educational system, to ensure quality, and to monitor progress. Without the validation evidence and systematic monitoring and evaluation of the accountability systems, the understanding whether goals are being met truthfully will not be known regardless how good the design of the accountability systems is and how valid the indicators are in representing the components of the accountability system.

The Center for Assessment has experience and successfully assisted states in designing an accountability system. For example, in recent years the Center for Assessment has successfully assisted Nevada, Utah, Wyoming, and Colorado in developing accountability systems. The Center for Assessment is considered a leader by the innovative demonstration authority and is currently working in Georgia, New Hampshire, US Virgin Islands, Utah, and Wyoming in developing their accountability systems compliant with ESSA and equally important representing the values and goals of each state. The Center for Assessment's operating principle and value is in to work closely collaborating with Arizona to design the accountability system. This is not only good practice, but is a requirement to ensure the accountability system both represents the values and goals of the Arizona and insures engagement, buy-in, and active participation by all stakeholders.

Additionally, the Center for Assessment has experience in evaluating large-scale initiatives utilizing qualitative and quantitative methods. So, one the components indicated above to ensure that the accountability system meets its intended goals and does so with fidelity involves evaluating the efficacy of the system. While few states have undertaken such efforts, it is important to understand and document the

impact and consequences of the accountability system. While this effort is a separate one from the development of the accountability system, the Center for Assessment has the expertise and is equipped to undertake this evaluation work.

2. Background

a. Provide a brief history of the organization and its governance structure, if applicable, or provide a brief overview of the individual's experience with accountability of K-12 public schools and districts
b. Identify the individuals from the organization that will be working with Arizona officials on all aspects of the accountability system's implementation.
c. Disclose and discuss the organization's work within and around Arizona's state education agency, local education agencies, and/or public education agencies including assurance any work or deliverables produced by the organization will exclude bias or undue influence, if applicable.

Response from the Center for Assessment:

The National Center for the Improvement of Educational Assessment, Inc. (The Center for Assessment) is a Dover, NH based not-for-profit corporation that seeks to improve the educational achievement of students by promoting improved practices in educational assessment and accountability. The Center for Assessment does this by providing services directly to states in conjunction with the states' large-scale assessment and accountability programs. The Center also works with organizations that work directly with states, or whose work impacts states, including the Council of Chief State School Officers (CCSSO), Achieve, The National Center for Educational Outcomes (NCEO) and the U.S. Department of Education. The Center also seeks to develop and disseminate broadly policies and practices that will improve educational assessment and accountability. The Center pursues the dissemination of best practices through an annual conference sponsored by the Center for Assessment; through extensive work with states' Technical Advisory Committees; through work with organizations that have extensive reach in areas of practical and policy advice, including CCSSO, NCEO, CRESST, Achieve; and through numerous publications and presentations at professional conferences.

The Center for Assessment currently has contracts with approximately 35 states/entities, several school districts, and many non-governmental organizations, and plays a significant role in multiple federal grants administered through state/federal research center partnerships. Since its inception in 1998, the Center was had contracts with approximately four-fifths of the states, and has worked with essentially all states in some capacities.

The National Center for the Improvement of Educational Assessment is a 501(c)(3) non-profit organization. Founded in September 1998, the Center's mission is to

improve the educational achievement of students by promoting improved practices in educational assessment and accountability. The Center focuses on the technical and practical issues that promote or inhibit the effectiveness of educational assessment programs. We seek to accomplish this mission by:

- Providing customized support to states and districts in designing, implementing, and improving fair, effective, and legally defensible assessment and accountability programs. The Center's staff provides a full range of support, including technical analyses, policy and management support, documentation and communication, and training. The Center also helps states design accountability systems that include effective programs in support of low-performing schools.
- Providing and managing Technical Advisory Committees that help ensure a state's evolving assessment and accountability programs receive the best on-going technical advice possible, focused on the specific issues and decision-making needs of the individual state or district.
- Developing and disseminating practical standards for assessment and accountability programs that include specific information about what states and districts should do today to have technically sound programs.
- Helping states develop innovative assessments, both standardized large-scale and comprehensive local assessment systems that feature integration with curriculum and instruction.
- Investigating and documenting at school, district or state levels strategies for educational improvement with promise of broader application.
- Advancing best practices in the field by serving as a conduit of information to all stakeholders in educational reform through sponsorship and leadership at conferences, the initiation of studies, and collaboration with other major service providers.

The Center for Assessment will include Scott Marion, Chris Domaleski, Damian Betenbenner, Juan D'Brot, and Thanos Patelis in serving, supporting, and actively working with Arizona in the development of the accountability system, facilitating and managing the advisory committees, undertaking any analysis and research efforts, writing and disseminating documents, and providing ongoing technical and informational assistance. The background, expertise, and experience for each follows.

Scott Marion, Ph.D. is the President and Executive Director of the National Center for the Improvement in Educational Assessment, Inc., a Dover, NH non-profit consulting firm. Scott Marion partners with Associate Director Chris Domaleski to manage the operations of the Center, and he works closely with the Center Board of Directors to establish the long- and short-term strategic direction of the organization. He is also actively engaged with Center clients; his projects include designing and supporting states in implementing assessment and accountability reforms, developing and implementing educator evaluation systems, and designing

and implementing high quality, locally-designed performance-based assessments. He is a national leader in designing innovative and comprehensive assessment systems to support instructional and accountability uses, including helping states and districts design systems of assessments for evaluating student learning of identified competencies.

Scott coordinates and/or serves on five or district state Technical Advisory Committees (TAC) for assessment, accountability and educator evaluation, including coordinating the PARCC assessment consortium TAC. He recently served on the National Research Council (NRC) committee responsible for designing a framework for next generation science assessments; he has also served on other recent NRC committees investigating the issues and challenges associated with incorporating value-added measures in educational accountability systems and on outlining best practices in state assessment systems.

Scott has published dozens of articles in peer-reviewed journals and edited volumes; he also regularly presents his work at the national conferences of the American Educational Research Association (AERA), National Council on Measurement in Education (NCME) and the Council of Chief State School Officers (CCSSO). In addition, Scott serves his community as Chair of the Rye (NH) School Board.

Prior to joining the Center for Assessment in early 2003, Scott was most recently the Director of Assessment and Accountability for the Wyoming Department of Education; he began his career as a field biologist and high school science teacher.

Scott received a Ph.D. from the University of Colorado Boulder with a concentration in Measurement and Evaluation.

Chris Domaleski, Ph.D. is Associate Director and partners with Executive Director Scott Marion to manage the operations of the Center. He also plays an active role as a consultant to multiple states supporting the development, implementation, and evaluation of assessment and accountability systems.

With a background in both psychometrics and policy, Chris advises education leaders in making sense of complex technical problems, and identifies real-world solutions to improve practice. Current interests and projects include designing innovative accountability systems that more fully incentivize and measure school quality, developing models for comprehensive assessment systems to support multiple purposes and uses, improving assessment design and practice for students with significant cognitive disabilities, and evaluating the effectiveness and impact of education policy.

He serves on several state technical advisory committees; is the coordinator of the Council of Chief State School Officers (CCSSO) State Collaborative on Accountability Systems and Reporting; is a technical advisor to two multi-state assessment consortia; and regularly provides technical support to the U.S. Department of Education. He also currently serves as an Associate Editor for the Journal of Educational Measurement, and regularly presents his research at national conferences.

Prior to joining the Center, Chris was Associate Superintendent for Assessment and Accountability at the Georgia Department of Education, where he was responsible for the development and administration of the state's K-12 testing program and accountability systems.

He received a Ph.D. from Georgia State University with a concentration in in Educational Policy Studies, concentrating in Research, Measurement, and Statistics and he has taught numerous graduate courses in measurement and statistics at Georgia State University and the University of Georgia.

Damian Betebenner, Ph.D. is a Senior Associate at the Center for Assessment. His work currently focuses on the development, implementation, integration and reporting/communication of state level growth analyses. He is the architect of the student growth percentile (SGP) model, which began in Colorado and has since been adopted in more than two-dozen other states. The National Council on Measurement in Education (NCME) recognized the model with its prestigious annual award for Outstanding Dissemination of Educational Measurement Concepts to the Public, given at the Annual Conference in May 2010 in Denver. Damian is also the primary architect for interactive Colorado Growth Model data visualization software that was recognized by Adobe Software as a Max Award Finalist at its 2009 Adobe Max convention for innovative uses of Adobe technology. Damian supports many state clients in implementing SGPs and applying SGP results to accountability determinations. Further, Damian conducts effective knowledge transfer of SGP implementation with his client states by supporting them as they eventually learn to implement the open-source SGP statistical package.

Prior to joining the Center for Assessment, he was an assistant professor in the Department of Educational Research Measurement and Evaluation, Lynch School of Education, Boston College.

Damian has earned two doctorates. He received a Ph.D. in Mathematics from the University of Wyoming and a Ph.D. in Educational Measurement from the University of Colorado, Boulder.

Juan D'Brot, Ph.D. is a Senior Associate at the Center for Assessment. Juan has led and contributed to work on developing ESSA-aligned accountability systems, growth models, exploring graduation options for students based on local legislation, peer review submissions, and revising readiness assessments for educational organizations and numerous states and jurisdictions. His work interests include assessment and accountability technical and policy issues, assessment and accountability design and implementation, measures of student growth, standard setting, educator accountability systems, and impact evaluation of policy and programs. Juan is especially interested in helping states and educational entities solve intricate problems in assessment and accountability design and implementation that often result from an intersection of policy, technical, and practical issues while navigating complex relationships between agencies.

Juan has been the author or co-author of various publications focusing on process and summative evaluations of supplemental educational services, teacher-focused professional development, the impact of accountability systems, the impact of interim assessment practices on summative assessment results, and the evaluation of technical assistance efforts provided to state and local education agencies throughout the country. He has also participated in over two-dozen invited or peer-reviewed presentations at professional conferences focusing on assessment, accountability, and research in education.

Prior to joining the Center, Juan was the Senior Director of Research at Data Recognition Corporation, where he provided leadership and assessment vision as the liaison between DRC Research and other departments to develop and disseminate strategic and technology-based solutions aligned to DRC's assessment programs. He was also responsible for designing, computing, and evaluating all traditional and IRT statistical analyses, including defining, managing, and monitoring all psychometric analyses. Previously, he served as the Director of Strategic Research Solutions for CTB and as the Executive Director of Assessment, Accountability, Research, and Evaluation for the state of West Virginia. As the Executive Director, he was responsible for the administration, development and implementation of all aspects of the statewide balanced assessment system, the state and federal accountability system, and providing strategic and direct oversight of grant-based and independent research and evaluation services for the department. During his tenure there, he implemented a balanced assessment system, transitioned the state to 100% online testing, implemented the West Virginia Growth Model, developed an approved growth-based accountability system under ESEA Flexibility, and led standard settings to define statewide cut scores for effective schools and teachers. He continues to leverage his previous experience as a research and evaluation specialist to help others understand the meaning behind quantitative and qualitative findings and to apply those results to policy and practice.

Juan received a Ph.D. from Capella University with a concentration in Industrial-Organizational Quantitative Psychology.

Thanos Patelis, Ph.D. is a Senior Associate at the Center for Assessment. Thanos has led efforts to evaluate the quality of assessments that include the design and implementation of psychometric analyses, validation studies, and a variety of statistical analyses. He has contributed to the development of methodologies for evaluating the quality of assessments. He has undertaken studies in evaluating the impact of large scale initiatives, accountability systems, and assessment programs. He has assisted in the design of performance assessments, non-cognitive assessments, score reports, growth models, and theories of action of educational initiatives. His areas of work are applied psychometrics, test validity, structural equation modeling, program evaluation, non-cognitive measurements, history of testing, growth modeling, school accountability systems and multivariate statistical analysis.

Prior to joining the Center for Assessment, Thanos was vice president of research and analysis at the College Board responsible for the evaluation of educational initiatives, statistical analysis of assessment data, performing linking studies with state and local assessments involving indicators of college readiness, supporting assessment development around non-cognitive skills, developing growth metrics and score reports and managing all data, policies, and procedures associated with research data. He also had developed and managed work plans and business functions of the research department including leading assessment score reporting products. Before his 15 years at the College Board, Thanos was a research associate at the Stamford (CT) Public Schools responsible for the testing program, evaluating programs, developing and providing in-service to teachers on classroom assessment, making presentations to educators and parents on testing results, and provided a variety of analysis and survey services.

Thanos has held leadership positions for regional, national, and international professional associations in educational measurements, educational research, and psychology. He is a fellow of the American Psychological Association, Division 5. He was received awards for his product development and measurement-related contributions, his teaching, and mentoring. He received his Ph.D. in psychometrics and master's in experimental psychology from Fordham University and his bachelor's in psychology from the College of the Holy Cross.

The Center for Assessment is in the process of developing a relationship for Damian Betenbenner to provide technical assistance in Arizona's growth model.

3. Overview of System

a. List and define the metrics included in a potential accountability system which meets the needs of Arizona's various school types and uses multiple measures. Please highlight the extent to which academic achievement on Arizona's statewide assessments (i.e. AzMERIT, NCSC, AIMS, etc.) and/or results from a menu of assessment can be meaningfully integrated in the proposed system.
b. Describe how the proposed metrics are aligned to college / career expectations and include any alignment studies, if available.
c. Describe how differentiated weights and metrics resulting in an overall letter grades can be compared between schools and across years to inform the following:
    i. Achievement of all students and progress of student subgroups
    ii. Information needed by parents/communities to inform school choice
    iii. Improvement of various types of schools within Arizona
    iv. Construct relevant components of school quality
d. Describe how the proposal may reduce administrative burden for LEAs and the SEA given the variation in accountability requirements related to applicable federal laws, state laws, State Board of Education rules, charter school authorizers, and other regulatory bodies.

Response from the Center for Assessment:

The request for information on the metrics of the accountability system cannot be prescribed or evaluated (as suggested by part c.iv) without a well-articulated theory of action where by the goals, purposes, and uses of the accountability system are specified (see Marion, 2010; Gong, 2008). Further, a description of how these metrics will affect the administrative burden for LEAs and the SEA cannot be provided without a full understanding of the current data collection efforts, infrastructure, and capacity. The Center for Assessment will accommodate the current data and information capacity and infrastructure as the Center for Assessment works with Arizona to develop the accountability system to meet the state's goals and values and ESSA's requirements.

The specification of the metrics cannot be done in a vacuum without the values and goals of Arizona forming the basis and the accommodation of current data infrastructure and capacity. In order for the metrics to be defined, the Center for Assessment recommends that first a leadership team, a steering committee of stakeholders, and a technical advisory committee be formed. The Center for Assessment will work with the leadership to team to hold a series of meetings with the leadership team and steering committee of stakeholders to (a) establish a theory of action about the accountability system as a means of establishing and documenting the goals, purposes and uses, (b) inform the requirements of ESSA and emerging regulations, (c) evaluate the current accountability system, and (d) gather input and feedback on the metrics to represent the components of the current accountability system.

The Center for Assessment has been successful in undertaking this approach in the development of accountability systems for states. These accountability systems represent the values and goals of states, received approval from the US Department of Education, and, have been characterized as innovative and useful.

It is important to note that while ESSA requires certain components of a state accountability system, there is flexibility in the specifics (Marion, 2016) and will be worked out in the rule making process (D'Brot, in press). The Center for Assessment has experience and successfully developed accountability systems that utilize state-specific and other available assessments and data as metrics.

The Center for Assessment has led the field with research to understand the technical requirements and necessary components of accountability systems (e.g., Betebenner, Diaz-Billelo, Domaleski, & Marion, 2014; Domaleski & Hall, 2014; Domaleski & Perie, 2012; Gong, Perie, & Dunn, 2006; Hill, Gong, Marion, DePascale, C., Dunn, J., & Simpson, M. A., 2005). Based on this experience, the Center for Assessment will provide the expertise needed in selecting the metrics and deciding on the manner of how to combine the components.

In order for this to be done, the Center for Assessment suggests that in addition to the process for (a) articulating the theory of action, (b) incorporating the ESSA law and regulations, and (c) ensuring feasibility of the collection and use of data, empirical evaluations of the decisions made be evaluated as part of the development effort. This aspect can and should be done by Arizona department resources, but the Center for Assessment is able and willing to assist in this analysis using data.

Some operating principles in the selection of the metrics can be offered, in addition to the driving forces in their selection of (a) alignment to the theory of action, (b) conforming with ESSA, and (c) feasibility, are suggested below:

1. Unless restricted by regulation, the use of dashboards of the metrics is a valuable approach in representing the metrics.
2. Compensatory models for combining measures have advantages that should be strongly considered.
3. Criterion-referenced approaches to evaluating acceptable performance on a metric are a good approach to use if there are clear policy determinations about what performance is valued.
    a. Criterion-referenced approaches ensure key equity outcomes can be prioritized.
4. Evaluation of metrics and design decisions to ensure the system does not systematically disadvantage any groups of students or schools
5. New metrics involving surveys require substantial development time to ensure technical quality is achieved.

The Center for Assessment is well equipped to assist Arizona in development the accountability and selecting metrics that are aligned to the goals of values of Arizona (once articulated), conform to ESSA requirements, and feasible. Metrics cannot be identified and the manner in which they are combined specified for use in an accountability system without engaging stakeholders and collaborating with Arizona staff members. In order for the accountability system to work (including the selection of metrics and methods for combining), the culture and context of data use (i.e., how they are collected, interpreted, and acted upon by communities of education, as well as officials) must be considered (Hargreaves and Braun, 2013) and doing so requires engaging the stakeholders in the process suggested earlier by the Center for Assessment.

4. Measuring Student Growth

a. What are the advantages of utilizing this measure of growth on Arizona's statewide assessments and in Arizona's new A-F Letter Grade Accountability System?
b. Please discuss evidence of technical appropriateness and statistical robustness to support the validity and reliability of student-level growth scores based on each of the following assessment scenarios:
    i. Vertically scaled assessments of grades 3 through 8 ELA standards
    ii. Vertically scaled assessments of grades 3 through 8 Mathematics standards
    iii. Vertically scaled, non-sequential with extreme variability in the instructional format for end of course assessments of high school ELA and Mathematics standards
    iv. Across test administration modality (Paper and computer-based) equated on a common vertical scale in each of the subjects above
    v. Varying assessments selected off of a menu of assessments potentially available in high school grades and administered in various modalities
    vi. Varying assessments selected off of a menu of assessments potentially available to students in elementary grades and administered in various modalities
    vii. Annual tests of English language proficiency as measured by AZELLA administered in Grades K through 12
    viii. Summative assessments of Science standards administered to students enrolled in grades 4, 8, and high school
c. Please describe the organization's professional experience and technical capacity for conducting this type of work on behalf of education agencies locally, nationally, and/or internationally.
d. Describe any services or assistance the vendor might provide to expedite the calculation of student growth scores so they are available to ADE, schools, students, and parents via student score reports produced by Arizona's test vendor(s).

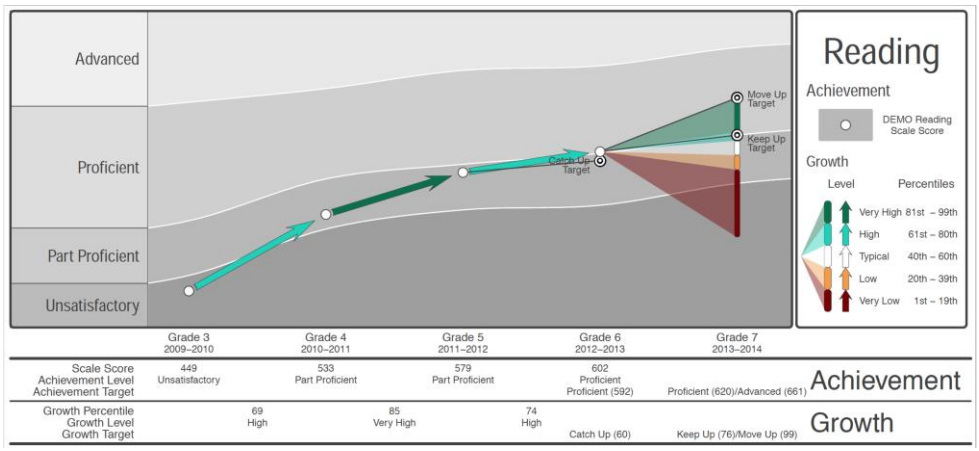
e. How can parents, teachers, students, and schools use growth score(s) to interpret individual student trajectory relative to Arizona's academic standards?
f. How can growth scores be aggregated and integrated into accountability determinations which may include varied weighting of proficiency results and other indicators of school performance?

Response from the Center for Assessment:

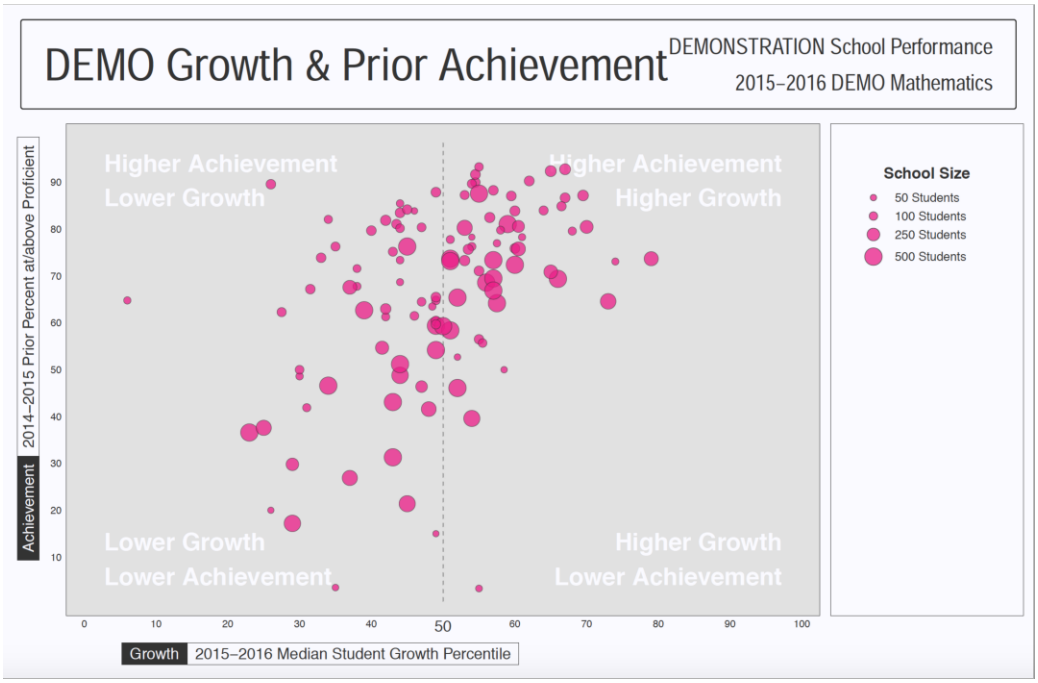
There are four views of school performance (see Carlson, 2001; Gong, 2002) that utilize both growth and status measures: Status, Status over time, individual student growth, individual student growth over time. What represents the most appropriate model will depend on the use that should be specified by the goals and values of the educational system (Goldschmidt, 2004; Gong, 2010). Based on earlier research on both status and growth models (Goldschmidt & Choi, 2007; Betebenner, 2008), the following recommendations are provided:

- Growth models using individual student growth provide the additional benefit of looking at student progress above and beyond student attainment/status. The additional information that growth models provide impacts stakeholder conversations in different ways:
    - At the individual student level, the addition of student growth data allows stakeholders to extend the conversation about how well a student did to include how much the student progressed. This extends conversations, particularly for low achieving students, to discuss situations where a student's attainment might not be what is desired but that their progress is remarkable and puts them on track to catch-up.

      To aid in helping communicate student level results, the Center for Assessment has worked extensively with over two dozen states including Colorado, Massachusetts, and Utah to produce visualizations to assist in communicating the simple conceptual message associated with student growth.



- At the school level with, for example, Arizona's statewide accountability system and in Arizona's Letter Grade A-F grading system, individual student growth provides a means of examining school quality using the additional characteristic of student growth: What is the level of attainment/status of students at a school versus how much are they learning/growing? The Center for Assessment has worked extensively with states to help communicate this distinction, often with the aid of visual representations like the "bubble plot" below.



The advantages of utilizing student growth as part of school accountability is that growth takes account of where a student starts so that one is discussing how much learning occurred, on average, at the school, which is often considered highly relevant in discussions of school quality. The SGP

methodology developed by the Center for Assessment allows states to interpret student growth (i.e., learning) in both norm- and criterion-referenced ways. That is, a norm-referenced interpretation allows users to understand how much learning occurred *relative to others* whereas a criterion-referenced interpretation allows users to understand how much learning occurred *relative to the performance standard* (i.e., proficiency or career and college readiness) the state has established.

Arizona has extensive experience with using individual student growth (SGPs) and we would recommend that continue to build/refine these efforts. These efforts would center on:

1. Effective/clear communication of growth results that go into the A-F calculations.
2. Investigation of both norm- and criterion-referenced components of growth as part of stakeholder values of what type of growth is most important.
3. Long-term investigation of establishing anchored growth-norms for purposes of investigating whether growth is increasing over time in the state.

- The quality and type of assessment can impact growth analyses. The Center for Assessment has extensive experience in understanding how the type of assessment impacts growth and the possible ways that growth analyses can be undermined due to assessment shortcomings.

  A prominent concept in the RFI relates to the issue of a vertical scale in measuring student growth. The issue of growth on vertically scaled tests comes up frequently, particularly of late, as many states now have tests that are vertically scaled. The existence of a vertical scale often gives stakeholders the impression that, "Growth is easy with a vertical scale". Our rejoinder to this often stated premise is that, "Subtraction isn't a growth model" (see https://view.literasee.io/Literasee/Georgia/report).

  A vertical scale is essential to understanding whether a student's score has increased/decreased from year-to-year, but is not suitable in and of itself for making determinations of whether a student's increase/decrease is exemplary or concerning (e.g., a 3 year old might grow 2 inches over the course of a year. Two inches is a well understood quantity but understanding whether 2 inches is good/bad requires one to go beyond subtraction).

  The RFI puts forward a number of specific testing scenarios and asks for a discussion of "evidence of technical appropriateness and statistical robustness to support the validity and reliability of student-level growth scores". We address these scenarios below.

- **Vertically scaled assessments of grades 3 through 8 ELA & Mathematics standards**.

  SGP calculations were originally developed in Colorado which had a vertical scale for its CSAP state assessment. SGP calculations, however, do not require a vertical scale. With the understanding the state assessments change, SGPs were developed to be as invariant to transformations of scale as possible. Thus, regardless of whether the testing system is based upon a consistent vertical scale year over year or changes to a different scale all together, SGPs allow for a common metric/vocabulary to be used to describe student progress.

  SGPs demonstrate moderate reliability/precision. For example, in most state analyses conducted/reviewed by the Center for Assessment, the standard error associated with an SGP range from 5 to 15 with a mean of 10. Taken in the context of student reporting, an SGP ranging for 1 to 35 (often called low) is highly likely to be less than a year's worth of growth, and SGP from 35 to 65 (often called typical) is likely to be a year's worth of growth, and an SGP from 65 to 99 (often called high) is likely to be more than a year's worth of growth.

  Due to measurement error in the tests utilize for the construction of SGPs, SGPs demonstrate slight bias. Measurement error adjustments are available via the SGP package using a SIMEX algorithm and are utilized by a number of states to create unbiased SGPs.

- **Vertically scaled, non-sequential with extreme variability in the instructional format for end of course assessments of high school ELA and Mathematics standards.**

  SGP analyses can be performed using non-sequential content area and/or grade progressions so long as there are a sufficient number of students from which to create a norm group. The Center for Assessment has extensive experience in this regard with multiple states, most prominently Georgia who administer over 10 end of course examinations at varying times during the year. In addition, if time varies extensively regarding when students are administered the exam, growth analyses can be performed with time and time-lag as independent variable. The SGP package currently includes this capability (SGPt analyses). For a review of the different types of course progressions performed please see configurations specified in Georgia EOC SGP Configurations

- **Across test administration modality (Paper and computer-based) equated on a common vertical scale in each of the subjects above**

  Multiple Center for Assessment clients have been impacted by test administration modality issues in the past two years as they transition to new, computer-based assessments. SGPs were utilized to determine, in fact, whether there were mode effects. In all cases there were mode administration effects. It is critical to understand that these mode administration effects are fundamentally about the status/attainment scores associated with the test. Often, because students are not randomly assigned to test administration mode, it is difficult to observe a mode effect. However, when SGPs are calculated the mode effect becomes clear.

  There are several options for dealing with mode effects:

  1. Create a mode adjustment at the scale-score level. That is, adjust the scores (usually the P&P scores) to conform to the online scale. This is the most comprehensive adjustment and will adjust status and growth scores accordingly. A few states that the Center currently works with are currently pursuing this option.
  2. Condition on administration mode within the growth analyses. Either running the growth analyses separately or including a dummy variable in the growth analyses indicating administration mode can achieve this option. This option deletes the mode effect from the growth analysis but does not change the status results. In addition, because students are not randomly assigned to administration mode, one or both groups of students may be unfairly (dis)advantaged by the adjustment since, by definition, the median SGP for each group will become 50.

- **Varying assessments selected off of a menu of assessments potentially available in high school grades and administered in various modalities.**

5. Other Technical Considerations

a. Describe any detailed analyses currently available or could be conducted to support the validity, reliability, and fairness of a proposed system as well as the methodology and validation process for standard setting the overall A-F letter grade determinations. The ADE will model all components for possible inclusion in a final accountability system.
b. Describe the timeline necessary to produce school accountability ratings which will differentiate and support the variety of public schools and districts within Arizona.

c. Describe the level of complexity and ability to replicate the statistical techniques which may be utilized throughout the system to differentiate school performance.
d. Describe any resources related to personnel, data, and/or technology the proposal may require, including any additional resources, data collection, management, and storage needed by the Department.
e. Please highlight any significant deviations from previous practice or changes to operational definitions currently utilized within Arizona's system of holding schools accountable.

The Center for Assessment has experience and expertise in undertaking evaluations of accountability systems that includes using a variety of methods. Because the process of validation involves an evaluative judgment that offers tentative conclusion based on partial evidence drawn from generally uncontrolled studies of schools and district (Braun, 2008), the process of validation involves a number of approaches, methodologies, and studies. It represents a series of efforts that should be undertaken by the responsible party.

The manner in which these are done is to build a chain of reasoning from the design and development process to the desired claims (interpretative argument) and gather theoretical and empirical support for the claims being made (Kane, 2001). Because accountability is different from assessments (Gong, 2008), the process involves social science research (Rossi, Lipsey & Freeman, 2004) and involves a variety of qualitative and quantitative approaches.

A full description of the proposed analyses and methodologies are beyond the scope of this response, because the effort involves understanding and incorporating (a) the goals of the program, (b) the context, (c) stakeholder perspectives, (d) program theory, (e) proposed uses of results, and (f) the scientific rigor (Patelis, 2012). However, following we present some key claims that should be investigated in the evaluation process along with exemplar studies to inform each. Although not comprehensive, these components are intended to capture the core areas that should be examined to evaluate the suitability of the model.

<u>Evidence Supports Claims in the Theory of Action</u>

This claim addresses the supports and structures that must be in place to bolster the integrity of the information in the model and to improve the likelihood that actions based on information derived from the accountability model will promote intended outcomes.

This broad claim connects to many aspects of Arizona's system including:

1. Assessments and indicators are reliable and valid given the appropriate context, purpose, and use

2. Academic growth information based on state and/or other assessments is credible and technically defensible.
3. Educators and leaders have access to the right information and have the knowledge, skills, and support necessary to improve student learning.

Results are Reliable

Reliability refers to the consistency or stability of a measure. In this case, we are interested in the reliability of the accountability indicators and outcomes.

There are multiple statistical approaches to evaluating the reliability of school or group determinations. However, at a minimum it is advisable to track the consistency of outcomes for various levels (e.g. schools, subgroups) within and across years. Although not without exception, it is expected that results will be well correlated for similar school types within year and for the same schools across years. Dramatic shifts in either classification of schools or characteristics of the distribution will signal a troubling lack of stability that will erode the credibility of the outcomes.

Results are Valid

If reliability addresses the extent to which the model provides a consistent answer, validity asks, "Is the answer correct?" Stated another way, to what extent are the results credible and useful for the intended purposes? At a minimum, an investigation of the validity of the model should address the following:

1. Is the model appropriately sensitive to differences in key factors?
2. Are the results associated with variables not related to effectiveness or generally those not under the control of the school, such as the socioeconomic status of the neighborhood?
3. Are the classifications credible?
4. Are negative consequences mitigated?

The first question addresses the extent to which the model differentiates outcomes among schools and/or classes. A model in which very few schools differ with respect to results (i.e. all ratings are high) will likely be out of sync with expectations and the credibility of the results will be suspect. Therefore, it is important to examine the distribution of results to determine if the outcomes are sensitive to differences and if the dispersion is regarded as reasonable and related to expected differences in school quality as documented from other means.

Second, it is important to examine the distribution of scores with respect to variables that should not be strongly associated with outcomes. For example, if there is a strong negative relationship between student poverty and school scores this suggests that effective schools are only those in which relatively affluent

students are enrolled. Such findings are implausible and erode credibility of the model.

The third question calls for examination of classifications with respect to external sources of evidence that should be correspondent with quality. For example, if the school accountability model is intended to identify and reward those schools that are preparing students for college and career, the validity evaluation will be incomplete without including data that reaches beyond K-12 and provides an indication of the post-secondary outcomes for graduates.

Finally, a validity evaluation should address the extent to which unintended negative consequences are mitigated. Some of these threats could be examined via survey data or focus groups, while others may be explored with extant data. Importantly, ongoing initiatives to gauge the extent to which positive outcomes outweigh potential negative side effects will bolster the consequential validity of this initiative and provide a mechanism to promote continuous improvement.

The Center for Assessment is excited to collaborate with the Arizona Department of Education to have these discussions to develop and, if needed, implement these methodologies to address these questions.

References

Arizona Department of Education (2015). *AZ Kids Can't Afford to Wait*. www.azed.gov/weheardyou

Betebenner, D., Diaz-Billelo, E., Marion, S., & Domaleski, C. (2014). Using student growth percentiles during the assessment transition: Technical, practical and political implications. Washington, DC: The Council of Chief State School Officers.

Braun, H. (2008). Viccissitudes of the validators. Presentation at the Reidy Interactive Lecture Series, Portsmouth, NH.

Carson, D. (2001/2006). Focusing state educational accountability systems: Four methods of judging school quality and progress. Dover, NH: The National Center for the Improvement of Educational Assessment. www.nciea.org

CCSSO (2011). Roadmap for next-generation state accountability principles. Washington, DC: Author.

Cour, K., Porter, W., Rome, A. M., & Towne, L. (2010). Promising practices in accountability: Report on the Chalkboard Project, Confederation of Oregon School Administrators, Oregon Business Association, and Stand for the Children. Seattle, WA: Education First Consulting. www.educationfirstconsulting.com

D'Brot, J. (in press). Every Student Succeeds Act: Considering the impact of college and career readiness indicators and design criteria for accountability systems. Dover, NH: The National Center for the Improvement of Educational Assessment. www.nciea.org

Domaleski, C. & Hall, E. (2014). Assessment transition and implications for accountability. Dover, NH: The National Center for the Improvement of Educational Assessment. www.nciea.org

Domaleski, C. & Perie, M. (2012). Promoting equity in state education accountability systems. Dover, NH: The National Center for the Improvement of Educational Assessment. www.nciea.org

Goldschmidt, P. (2004). Models for school accountability and program evaluation. Dover, NH: The National Center for the Improvement of Educational Assessment. www.nciea.org

Goldschmit, P. & Choi, K. (2007). The practical benefits of growth models for accountability and the limitations under NCLB. Policy Brief 9. Los Angeles, CA: UCLA, CSE/CRESST.

Gong, B. (2002). Designing school accountability systems. Washington, DC: CCSSO.

Gong, B. (2008). Validating Accountability Systems: Theory of Action. Dover, NH: The National Center for the Improvement of Educational Assessment. www.nciea.org

Gong, B. (2010). Using growth data to improve learning, teaching, and school functioning. Presentation at the CCSSO National Conference on Student Assessment, Detroit, MI.

Gong, B., Perie, M., & Dunn, J. (2006). Using longitudinal growth measures for school accountability under No Child Left Behind. Dover, NH: The National Center for the Improvement of Educational Assessment. www.nciea.org

Hargreaves, A. & Braun, H. (2013). Data-driven improvement and accountability. Boulder, CO: National Education Policy Center.

Hill, R., Gong, B., Marion, S., DePascale, C., Dunn, J., & Simpson, M. A. (2005). Using value tables to explicitly value student growth. Dover, NH: The National Center for the Improvement of Educational Assessment. www.nciea.org

Kane, M. T. (2001). Current concerns in validity theory. Journal of Educational Measurement, 38(4), 319-342.

Marion, S. F. (2016). Considerations for state leaders in the design of accountability systems under the *Every Student Succeeds Act*. Dover, NH: National Center for the Improvement of Educational Assessment. www.nciea.org

Peltzman, A. & Domaleski, C. (2010). Establishing a college- and career-ready accountability system in Washington state: Leveraging Washington's education reform plan. Washington, DC: Achieve, Inc.

Patelis, T. (2012). Ways not to get hurt when evaluating programs on violent behavior. Presentation at the annual convention of the American Psychological Association, Orlando, FL.

Rossi, P. H., Lipsey, M. W. & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7th Ed.). Thousand Oaks, CA: Sage Publications.

Tucker, M. S. (2014). Fixing our national accountability system. Washington, DC: The National Center on Education and the Economy. See www.ncee.org